SEAD Virtual Archive: Building a Federation of Institutional Repositories for Long-Term Data Preservation in Sustainability Science

Beth Plale,
School of Informatics and Computing,
Indiana University, Bloomington, Indiana, USA

Robert H. McDonald,
Data to Insight Center, IU Libraries,
Indiana University, Bloomington, Indiana, USA

Kavitha Chandrasekar,
Data to Insight Center,
Indiana University, Bloomington, Indiana, USA

Inna Kouper,
Data to Insight Center,
Indiana University, Bloomington, Indiana, USA

Stacy Konkiel,
IU Libraries,
Indiana University, Bloomington, Indiana, USA

Margaret L. Hedstrom,
School of Information,
University of Michigan, Ann Arbor, Michigan, USA

Jim Myers,
Computational Center for Nanotechnology Innovations,
Rensselaer Polytechnic Institute, Troy, New York, USA

Praveen Kumar,
Civil and Environmental Engineering,
University of Illinois, Urbana, Illinois, USA

The 8th International Digital Curation Conference will take place on 14–16 January 2013 in Amsterdam. Please ensure you use the guidance in this template to produce your paper. Please submit your paper in one of the following formats: Microsoft Word (.doc), Open Office (.odt) or Rich Text (.rtf).
http://www.dcc.ac.uk/events/idcc13/submissions

Abstract

Major research universities are grappling with their response to the deluge of scientific data emerging through research by their faculty. Many are looking to their libraries and the institutional repository as a solution. Scientific data introduces substantial challenges that the document-based institutional repository may not be suited to deal with. The Sustainable Environment - Actionable Data (SEAD) Virtual Archive specifically addresses the challenges of "long tail" scientific data. In this paper, we propose requirements, policy and architecture to support not only the preservation of scientific data today using institutional repositories, but also its rich access and use into the future.

Introduction

Major research universities are grappling with their response to the deluge of scientific data emerging as a result of their faculty's research. Many are looking to their libraries and institutional repositories as a solution. Research libraries have considerable expertise with preservation of the scholarly record. In cooperation with the IT organization of a university, a research library can play a significant role in preservation of large volumes of scientific data and in connecting the data to the published record of scholarship.

While the choice of using university resources for data storage may seem beneficial to the scientific community, the library, and the university, the general consensus is that institutional repositories are difficult to use. Often, it takes a long time and manual intervention to deposit data. If university institutional repositories are to compete with commercial and specialized scientific repositories, they need to provide services that are easy, fast, reliable, and community friendly. They need to satisfy the following requirements:

1. Ingesting of scientific data must be quick and minimally intrusive on a scientist's time.

2. Ingesting must be flexible enough to handle the varied kinds of data, i.e., collections of varied sizes, formats and composition.

3. Tools for advertising and serving data from an institutional repository need to be consistent with tools and processes of the scientific community.

The Sustainable Environment - Actionable Data (SEAD) Virtual Archive addresses the requirements listed above. It proposes policies and architecture to address the needs of sustainability science researchers, who study the physical, biochemical, and social interactions that affect our planet. Sustainability science is an example of long-tail science. Similar to Anderson's (2004, 2008) long tail of digital markets, the term characterizes a shift from a few highly demanded products (the head of the distribution curve) toward a large number of lower demanded "niche" products (its tail). Long-tail science consists of many research niches, which rely on customized methods and toolsets and on localized storage.

Sustainability science researchers often identify with disciplinary communities such as hydrology, ecology, or sociology; those communities have their own standards for data collection, description, and dissemination. The data are often designed to

satisfy the immediate needs of researchers, with less consideration for long-term availability and data reuse. The SEAD Virtual Archive (VA) supports not only the preservation of sustainability science data in the institutional repository today, but also aims to facilitate rich access and use in science and society into the future.

In the remainder of the paper, we discuss the design, policy decisions, and progress to date of the SEAD Virtual Archive. First, we give a brief overview of the SEAD Virtual Archive, situating it in the larger SEAD services framework as background. Next, we discuss processes and tools that must be in place to deposit scientific data into an institutional repository easily and with minimal effort. Then, we discuss the current architecture and planned approach to dealing with the heterogeneity of the long-tail sustainability data. We conclude by extending our discussion into the tools for advertising and serving data out from an institutional repository.

SEAD Virtual Archive Overview

The SEAD Virtual Archive is a federation layer over multiple institutional repositories. This layer offers the community of sustainability scientists a coherent view on their collective data. Even if institutional repositories removed major obstacles to data submission and researchers began to submit their data, the view of data would be a fragmented one; a researcher would have to search repositories one by one to find relevant data. The SEAD VA can provide a single view into the data for sustainability researchers. By leveraging inter-institutional cooperative agreements such as those that have been developed by the Committee on Institutional Cooperation (CIC), a federation layer such as the VA can form the basis of a cross-disciplinary problem-oriented resource.

SEAD VA software extends the open source software code developed by the Data Conservancy (Hanisch and Choudhury, 2009). Currently, the SEAD VA provides mechanisms to automatically deposit data into the Indiana University repository, *IUScholarWorks*[1], and the University of Illinois repository, *IDEALS*[2], with growth to other repositories.

Data arrives at the Virtual Archive from other SEAD services that capture metadata in RDF-based stores. These include an Active Content Repository (ACR), which stores data and metadata, and a social network repository, VIVO, which captures information about researchers and their projects, papers, and other scholarly products. Cross-reference metadata between these components also capture data-creator and data-publication connections. When a data set arrives in the Virtual Archive, it is presumed to be “ready to publish.” At this point, data and metadata are considered ready to be versioned/fixed (made immutable). We make the assumption at this phase of the project that the metadata also contains identified terms of access and use (i.e., repository licensing agreements are accepted). The processes initiated and acted upon the data submitted are captured in the Figure 1 below.

[1] http://scholarworks.iu.edu/repository/

[2] https://www.ideals.illinois.edu/

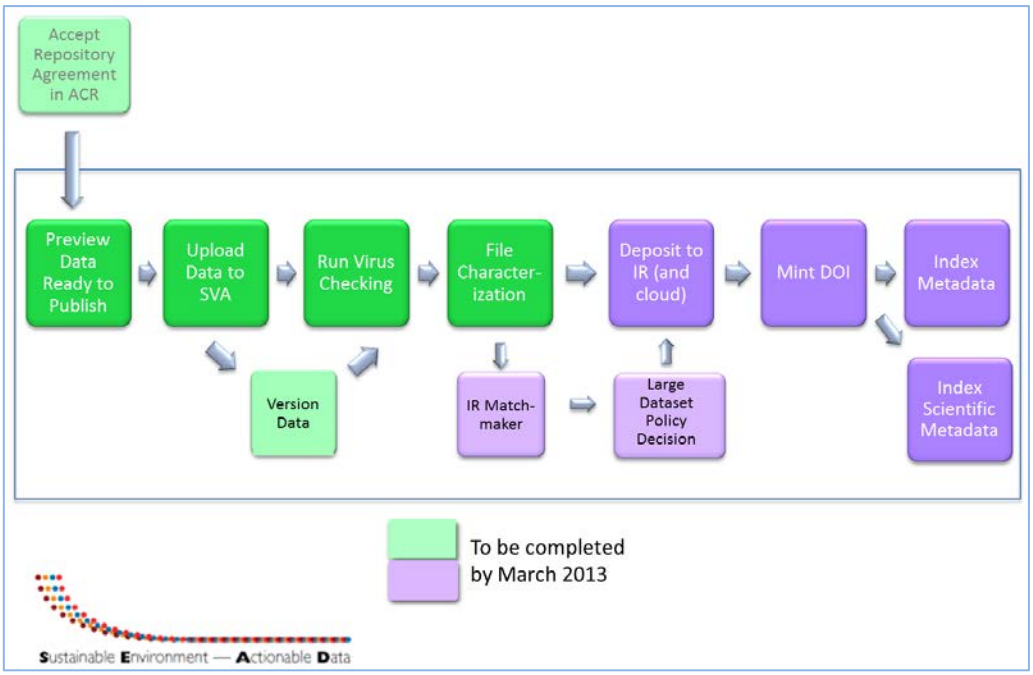

Figure 1. SEAD Virtual Archive workflow.

Query(ies) are then used to assemble metadata objects from any metadata that have been either entered manually, or harvested automatically by earlier services in the SEAD services framework. The metadata object also contains provenance about the file(s) to which the metadata is attached. SEAD VA transforms the RDF package into the SIP (Submission Information Package), an OAIS-compliant xml-encoded store for metadata.

Upon uploading datasets that were marked and selected for preservation, the SEAD VA checks dataset integrity and makes sure that the data are ready to deposit via its matchmaking mechanism. Matchmaking is a technical solution for automated deposit that reconciles the needs of institutional repositories with the needs of end users and the SEAD VA. For instance, a repository may only be able to take small files, or only want data from its affiliated researchers. These policy statements are encoded in the matchmaker, which evaluates the rules when deciding where to deposit a package. Because the SEAD VA interacts with several repositories that all have different scopes, missions, service orientations and depositor requirements, we need to make sure that such reconciliation satisfies all partners.

Easing the Burden of Data Submission

Many processes, policies, and tools need to be in place in order to make the ingest of scientific data into an institutional repository easy and smooth. The SEAD Virtual Archive has implemented a SWORD API client for automated data deposit into the repositories. SWORD is a widely accepted deposit protocol and can be used for a wide range of deposits (Stuart, Castro, Jones, 2012). Our plan is to extend the SWORD API deposit to non-DSpace repositories (for example, repositories based on

Fedora software). We have also implemented a separate automated service for the creation of communities and collections for *IUScholarWorks* and *IDEALS*.

Licensing Agreements

The SEAD project aims to follow the best practices for providing clear guidance to users of data in order to alleviate the complexities and ambiguities surrounding data release (Ball, 2012). We are engaged with legal counsel from several institutions involved in the project on developing clear statements that would a) explain the rights and responsibilities associated with the use of datasets and b) direct users to relevant information regarding licensing. Such statements will encourage users to assign licenses to their datasets and make the process of licensing data easier.

In addition to allowing users to have control over how their data will be used, the SEAD VA recognizes the needs of institutional repositories to have the rights to distribute data and perform necessary transformations for long-term preservation (e.g., migrating data to more suitable formats or changing data to protect human subjects).

Repositories set their terms in distribution licenses, which depositors must agree to when they make submissions. Such licenses usually include language that grants the repository basic rights to store and re-format files that help ensure that the repository can function in perpetuity. Once prepared for deposit, datasets usually cannot be modified by either library staff or data creators. Some repositories, though, are different. The ICPSR archive from University of Michigan, for example, retains the right to make changes to data to remove sensitive information in order to protect human subjects. Some repositories also retain the right to make metadata public and require confirmation of compliance with sponsoring organizations, while others do not contain such clauses in their licensing agreements.

To ensure that we address the rights of both depositors and repositories, the SEAD Virtual Archive team is developing two options in parallel: a single-license solution that would satisfy all repositories and a programmatic service that can identify when there is a match between the independently stated requirements of specific end users, repositories, and SEAD Virtual Archive.

Persistent Identification

The SEAD project applies persistent identifiers (PIDs) to datasets and authors to track usage and citations. For dataset identification the SEAD project uses DOIs that are being resolved to the institutional repositories that have accepted these research datasets for publication. For author identification, we currently use internal VIVO identifiers, which are unique links to each researcher profile.

Dataset Identification: Handle versus DOI

DOI is based on the Handle system, which is an established standard for PID resolution supported by the DSpace institutional repository software. Many repository support staff are already familiar with Handles and there are over 1300 installations world-wide. The decision to use the DOI system in SEAD on top of the Handle

system was made based on concerns over limitations of the Handle system, as well by the desire for the SEAD VA to work with multiple repositories. Handle IDs do not store associated metadata, while DOIs do—an important concern for a project developing rich metadata and an additional preservation measure. Further, while Handles are generated primarily for use in institutional repositories, DOIs are supported by a wide range of publishers and vendors as the citation standard and are used widely for supplying a PID for newly created research outputs. Using DOIs therefore allows SEAD VA to go beyond DSpace-based repositories, to federate across a larger number and variety of repositories, and to integrate with other data preservation and citation technologies.

DOIs are minted by a DOI creation service in the SEAD VA workflow that utilizes the DataCite API (EZID). The DOI information is then added as metadata to the files deposited in IRs. After the files are deposited, new metadata for the DOI are transmitted to EZID, including the new DOI target URL that is now the IR URL for the data. Among the other metadata transmitted to EZID is location (URL), title, author, and publication year.

One of the limitations for the DOI standard is that even though it allows for multiple destination resolution[3], most implementations do not support it. Ideally, the DOI would resolve to the data set's locations in both the VA and ACR components of SEAD, so end users could access both the active and the fixed data version. Currently, DataCite does not allow this action, so we overcome this limitation by maintaining a DOI and a separate link to the ACR in both the SEAD VA and SEAD-VIVO.

DOIs in SEAD VA are assigned at the top collection level. Assigning DOIs at collection level supports data citation by preserving a pointer to the entire collection. Separate DOIs are also assigned for each item that contains file(s) within the collection. Generating DOIs for items (folders) will help cite subsets of files within a collection. These generated DOIs are stored as item metadata during the ingestion process.

Author identification: VIVO ID versus ORCID ID

The SEAD VA needs clear mechanisms for user identification and rights managements based on their affiliation and repositories' preferences. Fortunately, a core service in the SEAD services framework is the SEAD-VIVO, a social networking tool that connects researchers with their publications and datasets.

Currently, the SEAD project uses the built-in VIVO unique ID for all researcher profile information stored in our Active/Social Content Repository. This VIVO ID works much like the Handle system described above and can be effectively used primarily at the domain or institutional level[4]. However, the VIVO ontology (Mitchell et al, 2011) currently supports many different researcher PID system IDs, including

[3] See the DOI handbook, section 3.3 Multiple resolution http://www.doi.org/doi_handbook/3_Resolution.html#3.3

[4] ORCID Non-profit Community [http://about.orcid.org/about/what-is-orcid]

ORCID[5], ResearcherID (Thomson-Reuters), Scopus Author ID (Elsevier), Pivot ID (Proquest), and others.

Now that the ORCID technical infrastructure has been established by the ORCID non-profit researcher ID community, SEAD hopes to take advantage of that service by enabling ORCID ID registration as the main unique identifier system used with SEAD. This would enable researcher identification at a more global level and offer many of the benefits that have been described here in our section on DOI support as a PID for data set publication.

In our initial work with ORCID, we aim at answering the following questions within our prototype test-bed:

1. If SEAD uses ORCID as a researcher ID service for SEAD-affiliated researchers, can SEAD as a data curation and preservation organization, take advantage of activity reporting and updates culled from ORCID to provide more up-to-date profile content for researchers?

2. What is the best way to show productivity or activity outcomes from researchers involved in SEAD data publishing and for what audiences?

3. Can we enable a light-weight, third-party authentication mechanism for SEAD that utilizes the ORCID PID as the authentication mechanism for SEAD researchers?

4. What processes and mechanisms need to be in place to tie more closely ORCID PIDs with the SEAD VA?

Handling the Heterogeneity of Data

As it was pointed out above, sustainability science is broad, and the data are diverse. We are working with domain scientists in order to map this diversity into a set of flexible yet manageable categories. So far, we have been working with data sets that are diverse in their size, formats, and structure. We have identified the following five canonical cases for data preservation and discovery in sustainability science that the SEAD VA must support:

1. Small-sized collection (overall size is less than 150 Mb).

2. Large-sized collection (overall size is more than 1 Gb).

3. Heterogeneous collection of related files with flat or hierarchical structure.

4. Collection that contains many interconnected variables (a relational database).

[5] http://orcid.org/

5. Collection that contains formats that are unique to the subdiscipline or community.

The SEAD Virtual Archive offers automated capabilities to handle the first case, i.e., the ingest of small-sized collections that do not contain unique file formats. Known formats make the ingest process unproblematic. SIPs are generated in the SEAD VA and prepared for storage and management, along with structural and descriptive information about the contents of files.

Handling large datasets is an issue that requires cooperation among the repositories, IT units, and the SEAD VA team. Institutional repositories are often deployed with limited resources; for example, they run on machines that can handle only a certain amount of data. We are looking into solutions that would allow the SEAD VA to overcome such limitations. One solution involves storing files in the cloud or in other university storage facilities, such as the High Performance Storage System (HPSS) developed through the university consortium and a partnership between US government labs and the IBM Corporation. The challenge for the SEAD VA is to automatically create links to entities in such storages and register those links in the repositories.

Preserving heterogeneous collections involves forming SIPs based on the large numbers of files, where each file may have not only its own structural metadata, i.e., information about size, format, and so on, but also its own scientific metadata. For structural metadata the SEAD VA adapted the OAIS model and its packaging framework. For heterogeneous scientific metadata, we are looking into indexing solutions, such as the use of open-source search platform Solr[6] or stand-alone metadata catalogs. Such solutions will provide upper level integration of all types of metadata and allow users to search and retrieve subsets of heterogeneous collections.

The problem of data sets that contain many interconnected variables or that are stored in a relational database is that such data cannot be deposited as a .zip or binary file, because then their structural, descriptive and scientific metadata may be lost. To be usable, databases first need to be transformed into a secondary form that would allow for rich metadata to be preserved. We are working on the solution to transform relational databases into an XML-compliant store that would contain information about variables so that they can be queried and searched.

The last case, handing unique or unknown formats, is at the very early stage of development in the SEAD VA. Such formats are treated as binary files, i.e., no additional information about their format and structure is extracted. As a result, no extensive preservation services, such as migration services, can be guaranteed for such formats at this point. One possible solution that is under investigation is to query external file format services and extract relevant information from them. Another solution is to push data sets back to data creators and request additional metadata that could be entered manually via a metadata editor or extracted by running other metadata extraction services.

[6] http://lucene.apache.org/solr/

Conclusion

In this paper we described the SEAD VA as a single federated layer over multiple repositories. We discussed the key issues in making institutional repositories play the major role in data services for sustainability science. We identified key requirements driven by this long-tail science community, and reported on our decisions and plans.

The SEAD VA as a federated solution has the benefit of giving a unified view of a sustainability science data resource even though the data are drawn from multiple institutional repositories. In addition to serving as a federated deposit service, the SEAD VA performs another crucial function in data services. One could envision a repository supporting multiple lightweight federation services like SEAD VA, each of which serves a particular scientific community. If each federation service supports its outward interfaces via record-exposing protocols such as the ones developed by the Open Geospatial Consortium (OGC) or DataONE, a network of federated services would become a scientific search portal with rich discovery interface. Such a portal will track and index information from institutional repositories and expand their services, providing more breadth and accuracy, and at a less cost to the libraries.

Acknowledgements

This work funded by the National Science Foundation under cooperative agreement #OCI0940824 and by the grant from the Council on Library and Information Resources and the Alfred P. Sloan Foundation, award #4112440.

References

[online magazine] Anderson, C. (2004). The Long Tail. Wired, 12.10. Retrieved from http://www.wired.com/wired/archive/12.10/tail.html

[book] Anderson, C. (2008). The Long Tail, Revised and Updated Edition: Why the Future of Business is Selling Less of More. New York, NY: Hyperion Books.

[report] Ball, A. (2012). 'How to License Research Data'. *DCC How-to Guides.* Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/how-guides

[unpublished proceedings] Hanisch, R. & Choudhury, S. (2009). The Data Conservancy: Building a Sustainable System for Interdisciplinary Scientific Data Curation and Preservation. Paper presented at the PV 2009 conference, Madrid, Spain. Retrieved from http://jhir.library.jhu.edu/handle/1774.2/34018

[conference paper] Mitchell, S, S. Chen, M. Ahmed, B. Lowe, P. Marks, N. Rejack, J. Corson-Rikert, B. He, and Y. Ding. (2011). The VIVO Ontology: Enabling Networking of Scientists. *ACM WebScience Conference*, Koblenz, GA, June 14-17.

[journal article] Stuart, L., Castro, de P., & Jones, R. (2012). SWORD: Facilitating Deposit Scenarios. D-Lib Magazine, 18(1/2). Retrieved from http://www.dlib.org/dlib/january12/lewis/01lewis.html